Spatial genomic heterogeneity within localized, multifocal prostate cancer

nature genetics

Spatial genomic heterogeneity within localized, multifocal prostate cancer

Paul C Boutros[1–3], Michael Fraser[4,23], Nicholas J Harding[1,23], Richard de Borja[1,23], Dominique Trudel[5,23], Emilie Lalonde[1,2], Alice Meng[3], Pablo H Hennings-Yeomans[1], Andrew McPherson[6], Veronica Y Sabelnykova[1], Amin Zia[1], Natalie S Fox[1,2], Julie Livingstone[1], Yu-Jia Shiah[1], Jianxin Wang[1], Timothy A Beck[1], Cherry L Have[5], Taryne Chong[1], Michelle Sam[1], Jeremy Johns[1], Lee Timms[1], Nicholas Buchner[1], Ada Wong[1], John D Watson[1], Trent T Simmons[1], Christine P'ng[1], Gaetano Zafarana[4], Francis Nguyen[1], Xuemei Luo[1], Kenneth C Chu[1], Stephenie D Prokopec[1], Jenna Sykes[7], Alan Dal Pra[8], Alejandro Berlin[8], Andrew Brown[1], Michelle A Chan-Seng-Yue[1], Fouad Yousif[1], Robert E Denroche[1], Lauren C Chong[1], Gregory M Chen[1], Esther Jung[1], Clement Fung[1], Maud H W Starmans[1], Hanbo Chen[1], Shaylan K Govind[1], James Hawley[1], Alister D'Costa[1], Melania Pintilie[7], Daryl Waggott[1], Faraz Hach[6], Philippe Lambin[9], Lakshmi B Muthuswamy[1], Colin Cooper[10–12], Rosalind Eeles[10,13], David Neal[14,15], Bernard Tetu[16], Cenk Sahinalp[6], Lincoln D Stein[1], Neil Fleshner[17], Sohrab P Shah[18–20], Colin C Collins[21,22], Thomas J Hudson[1], John D McPherson[1], Theodorus van der Kwast[5] & Robert G Bristow[2,4,8]

Herein we provide a detailed molecular analysis of the spatial heterogeneity of clinically localized, multifocal prostate cancer to delineate new oncogenes or tumor suppressors. We initially determined the copy number aberration (CNA) profiles of 74 patients with index tumors of Gleason score 7. Of these, 5 patients were subjected to whole-genome sequencing using DNA quantities achievable in diagnostic biopsies, with detailed spatial sampling of 23 distinct tumor regions to assess intraprostatic heterogeneity in focal genomics. Multifocal tumors are highly heterogeneous for single-nucleotide variants (SNVs), CNAs and genomic rearrangements. We identified and validated a new recurrent amplification of *MYCL*, which is associated with *TP53* deletion and unique profiles of DNA damage and transcriptional dysregulation. Moreover, we demonstrate divergent tumor evolution in multifocal cancer and, in some cases, tumors of independent clonal origin. These data represent the first systematic relation of intraprostatic genomic heterogeneity to predicted clinical outcome and inform the development of novel biomarkers that reflect individual prognosis.

Prostate cancer is the most commonly diagnosed male malignancy in developed countries[1]. Although the majority of prostate cancer is diagnosed as organ-confined disease, cancers with similar Gleason scores (for example, Gleason scores 7–10) show substantial interpatient heterogeneity and differential prostate cancer–specific mortality rates[2,3]. A further complexity lies in intraglandular biological heterogeneity between individual cancer foci; indeed, ~80% of prostate cancers contain >1 disease focus[4]. Local therapy fails in up to 30–40% of patients despite the presence of homogeneous clinical risk parameters (the same NCCN (National Comprehensive Cancer Network) risk category based on similar TNM stages, Gleason scores and pretreatment PSA (prostate-specific antigen) values)[2,3,5]. Therefore, the genomic interrogation of prostatic lesions within and between patients could identify different pathways of tumor

[1]Ontario Institute for Cancer Research, Toronto, Ontario, Canada. [2]Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada. [3]Department of Pharmacology and Toxicology, University of Toronto, Toronto, Ontario, Canada. [4]Ontario Cancer Institute, Princess Margaret Cancer Centre, University Health Network, Toronto, Ontario, Canada. [5]Department of Pathology and Laboratory Medicine, Toronto General Hospital, University Health Network, Toronto, Ontario, Canada. [6]School of Computing Science, Simon Fraser University, Burnaby, British Columbia, Canada. [7]Department of Biostatistics, Princess Margaret Cancer Centre, University Health Network, Toronto, Ontario, Canada. [8]Department of Radiation Oncology, University of Toronto, Toronto, Ontario, Canada. [9]Department of Radiotherapy, Maastricht University, Maastricht, the Netherlands. [10]Division of Genetics and Epidemiology, The Institute of Cancer Research, Sutton, UK. [11]Department of Biological Sciences, University of East Anglia, Norwich, UK. [12]School of Medicine, University of East Anglia, Norwich, UK. [13]Royal Marsden National Health Service (NHS) Foundation Trust, London and Sutton, UK. [14]Urological Research Laboratory, Cancer Research UK Cambridge Research Institute, Cambridge, UK. [15]Department of Surgical Oncology, University of Cambridge, Addenbrooke's Hospital, Cambridge, UK. [16]Department of Pathology, Laval University, Quebec City, Quebec, Canada. [17]Division of Urology, Princess Margaret Cancer Centre, University Health Network, Toronto, Ontario, Canada. [18]Department of Pathology, University of British Columbia, Vancouver, British Columbia, Canada. [19]Department of Computer Science, University of British Columbia, Vancouver, British Columbia, Canada. [20]British Columbia Cancer Agency Research Centre, Vancouver, British Columbia, Canada. [21]Department of Urologic Sciences, University of British Columbia, Vancouver, British Columbia, Canada. [22]Laboratory for Advanced Genome Analysis, Vancouver Prostate Centre, Vancouver, British Columbia, Canada. [23]These authors contributed equally to this work. Correspondence should be addressed to P.C.B. (paul.boutros@oicr.on.ca) or R.G.B. (rob.bristow@rmp.uhn.on.ca).

Received 22 December 2014; accepted 1 May 2015; published online 25 May 2015; doi:10.1038/ng.3315

progression and lead to bespoke prognostic genomic information to use in stratified treatment protocols[4].

Biomarkers based on CNAs or mRNA abundance in primary tumor[6–10] or blood[11,12] samples have not yet reached maximal clinical application owing to a lack of understanding of inherent intraglandular and multifocal heterogeneity[13]. Individual prostate cancer foci are believed to be clonal, but their molecular nature within a given patient remains largely uncharacterized at whole-genome resolution[14]. Moreover, the potential impact of intraprostatic heterogeneity with respect to known prognostic genomic aberrations has not been studied. This information would complement whole-genome and exome sequencing data on recurrent mutations in metastatic castration-resistant disease (mCRPC)[15–21] by providing information on the relative aggressiveness of foci with similar Gleason scores. Thus far, the interfocal heterogeneity of localized prostate cancer within a given prostate gland has not been explored, despite such knowledge being critical in personalizing patient treatment with genome-based biomarkers. Herein we comprehensively annotate the unique genomic characteristics of 23 individual prostate cancer foci from 5 patients and describe, for the first time to our knowledge, a subset of *MYCL*-associated cancers, suggesting unique pathways involved in prostate cancer progression. These studies provide genomic portraits of the intra- and intertumoral molecular landscape of multifocal, potentially curable localized prostate cancer in the context of developing bespoke treatment options.

RESULTS

Genomic landscape of potentially curable prostate cancer

To explore interpatient heterogeneity within a homogeneous prognostic risk category (NCCN intermediate risk), we examined the genomic profiles of 74 prostate cancer specimens derived from patients with Gleason score 7 pathology (consisting of 57 pretreatment biopsies and 17 surgically resected prostates) (**Fig. 1a**). Samples with >70% tumor cellularity from the index lesion (defined as the largest focus of disease) underwent genome-wide copy number analysis using Affymetrix OncoScan. CNA analyses found extensively heterogeneous profiles (**Fig. 1b** and **Supplementary Table 1**) associated with highly variable percentages of genome alteration (PGA; median tumor PGA of 4.43%, encompassing 2,559 genes; range of 0–16.3%, encompassing 0–10,133 genes).

Gleason score 7 tumors (primary grade + secondary grade: 3 + 4 and 4 + 3) showed similar diversity in genomic instability and specific CNAs (**Fig. 1b**), with statistically indistinguishable PGA values (median of 3.56% versus 4.77%; $P = 0.69$) and number of genes altered (median of 3,071 versus 2,438 genes; $P = 0.55$). Notably, no specific genes were altered at different frequencies between the Gleason score 3 + 4 and 4 + 3 subsets (**Supplementary Fig. 1**), and neither clinical T category (**Supplementary Fig. 2**) nor pretreatment PSA value (**Supplementary Fig. 3**) was associated with specific CNAs in these prostate cancers of similar pathology.

Recurrently altered genes in Gleason score 7 prostate cancer

Notably, we identified genomic abnormalities previously associated with locally aggressive cancer or mCRPC in a subset of the localized Gleason score 7 tumors. Chromosome 8p, harboring the tumor suppressor *NKX3-1*, was deleted in 4 + 3 (8/28) and 3 + 4 (15/46) tumors. Similarly, chromosome 8q, containing the *MYC* oncogene, was amplified in 4 + 3 (6/28) and 3 + 4 (5/46) tumors. We identified 36 recurrent focal CNAs comprising 115 genes ($q < 0.01$; **Supplementary Tables 2** and **3**), including several cancer-associated genes. The most significantly deleted locus was 17p13.1, containing *TP53* ($q = 6.56 \times 10^{-10}$). *EGFR* was the only gene in a recurrently amplified locus at 7p11.2 ($q = 0.000127$). However, we also identified aberrations in several genes not previously implicated in localized prostate cancer; most interesting of these was a CNA affecting the MYC family member L-MYC, encoded by *MYCL* at 1p34.2 ($q = 1.22 \times 10^{-31}$). This was a highly focal amplification, with a minimally amplified region of 8.5 kb (median of 22 kb; range of 8.5 kb to 1 Mb) encompassing the *MYCL* 5′ UTR and coding region (**Supplementary Fig. 4**). We detected *MYCL* amplification in both Gleason score 3 + 4 (13/46) and 4 + 3 (7/28) tumors at similar frequencies ($P = 0.86$), and this amplification was more common than *MYC* amplification (**Fig. 2a**) in this cohort. MYCL gain in prostate cancer has only been reported in a mouse model of p53- and Rb-induced prostate cancer[22] and at a low frequency in a small cohort of hormone-refractory patients[23].

We validated this unexpected finding by quantitative RT-PCR (qPCR) with probes within the 8.5-kb minimally amplified region, using the NCI-H510A non–small cell lung cancer cell line as a positive control; qPCR validated *MYCL* copy number status in >96% of the 169 distinct specimens tested (59/64 with amplification and 105/105 copy number neutral; **Supplementary Table 4**). We detected *MYCL* amplification in samples from 2 different hospitals in separate provinces (32/115 samples from Toronto and 27/54 samples from Quebec City) and in both biopsy specimens (16/70) and radical prostatectomies (48/169). We further validated the focal nature of the *MYCL* amplification using ten additional qPCR probes flanking (within ~1 Mb of) the minimally amplified region (**Supplementary Fig. 5**); no specimens with validated *MYCL* amplification showed concurrent amplification of flanking regions at 1p34.2-3, whereas a control cell line (NCI-H510A) showed amplification with all ten probes, as expected (**Supplementary Table 5**). Finally, FISH analysis (**Supplementary Fig. 6**) of 5 prostate cancers with PCR-validated *MYCL* amplification showed >2 copies of the gene in 8–20% of glands, supporting the hypothesis that *MYCL* amplification is heterogeneous within an individual prostate. In contrast, we observed no evidence of *MYCL* amplification in prostate cancers with PCR-validated copy-neutral *MYCL* or in four benign prostates from men who underwent cystoprostatectomy (with no pathological evidence of prostate cancer). These data strongly support the presence of a focal amplification of *MYCL* in a subset of localized prostate cancers.

We then investigated the copy number status of all three *MYC* family members by microarray, including the *MYCN* isoform associated with neuroblastomas and retinoblastomas[24,25]. At least one *MYC* isoform was amplified in 39 of 75 patients, and these aberrations were largely mutually exclusive. We observed concurrent amplification of *MYC* and *MYCL* in only one patient ($P = 0.066$), consistent with MYC autorepression[26]. Amplification of any *MYC* isoform was associated with increased genomic instability (**Fig. 2b** and **Supplementary Table 6**). Whereas *MYC*-amplified tumors infrequently exhibited loss of *TP53* (2/13 patients), *MYCL*-amplified tumors almost always did (19/21 patients; $P = 5.93 \times 10^{-5}$, Fisher's exact test), indicating that dysregulation of different MYC family members leads to distinct molecular consequences. *MYC*-amplified tumors had marginally higher PGA values than *MYCL*-amplified tumors (7.61% versus 5.88%; $P = 0.14$; **Fig. 2c**), but *MYCL*-amplified tumors harbored a larger number of smaller aberrations (333 versus 7; $P = 2.10 \times 10^{-5}$). We compared the copy number profiles of *MYC*- and *MYCL*-amplified tumors and identified 1,438 genes that showed different copy number frequencies ($q < 0.05$) in these two groups (**Supplementary Table 7**). We confirmed these results at the mRNA level with mRNA abundance profiling of 24 samples and demonstrated that these changes were not secondary to altered *TP53*, as tumors showing only *TP53*

deletion (but not concomitant *MYCL* amplification) had globally different CNA and mRNA abundance profiles than tumors also harboring *MYCL* amplification (**Supplementary Fig. 7**). In total, 294 genes ($q < 0.05$) showed different mRNA abundance in *MYCL*-amplified tumors than in tumors that did not harbor amplification of any *MYC* family member. *MYCL*-amplified tumors showed alterations of genes involved in desmosome assembly as well as ARF GTPase activation (**Supplementary Table 8**). Similarly, despite the small numbers of

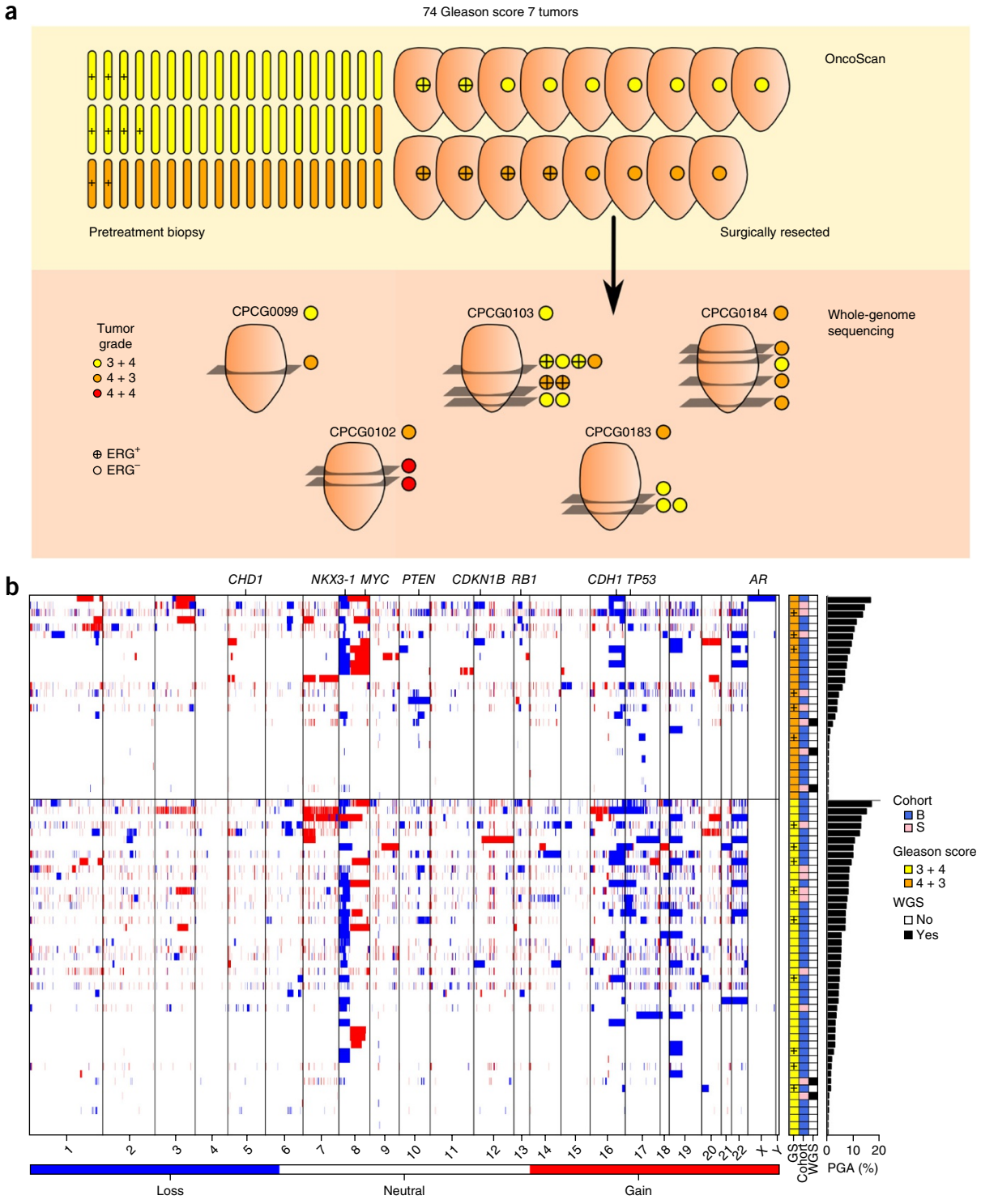

Figure 1 The landscape of Gleason score 7 prostate cancer. (**a**) We initially surveyed 74 Gleason score 7 prostate cancers consisting of 57 pretreatment biopsies and 17 surgically resected prostates. All 74 samples were processed using Affymetrix OncoScan microarrays, and 5 surgically resected samples were selected for whole-genome sequencing. To investigate the multifocality of the surgically resected tumors, multiple regions of the prostate were used to identify and compare genomic abnormalities. (**b**) Genome-wide view of CNAs in 74 patients, each annotated with Gleason score (GS), treatment modality (B, pretreatment biopsy; S, surgically resected) and whether whole-genome sequencing (WGS) was performed. Patients are sorted by PGA from highest to lowest, with values ranging from almost 20 to 0.0% (histogram to right).

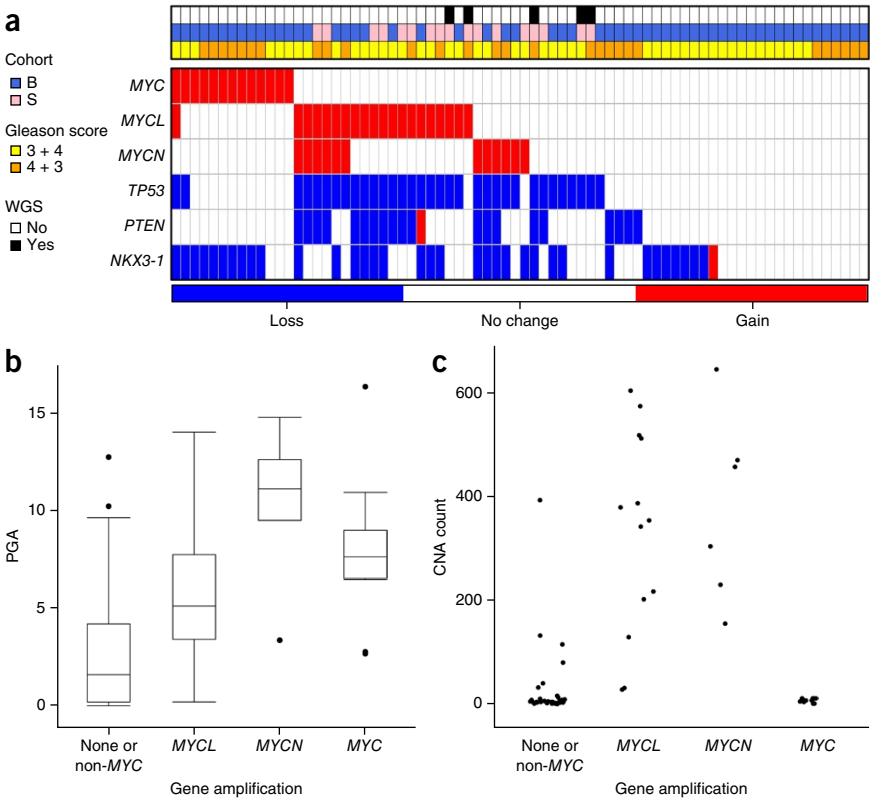

Figure 2 Recurrent *MYCL* mutations. MYC family members are altered in prostate cancer. GISTIC analysis showed *MYCL* to be recurrently mutated in Gleason score 7 tumors. (**a**) *MYCL* amplification exhibits strong co-occurrence and mutual exclusivity patterns with other major prostate cancer genes, including *MYCN* and *MYC* (simulation analysis, $P = 0.0011$; Online Methods). (**b**) PGA is increased with amplification of any *MYC* isoform (*t* test in comparison to samples with no *MYC* amplification: $P_{MYCL} = 0.018$, $P_{MYCN} = 3.8 \times 10^{-3}$, $P_{MYC} = 3.7 \times 10^{-4}$. Tukey box plots are shown. (**c**) The number of CNAs varies among patients displaying amplification of *MYC* family members (*t* test in comparison to samples with *MYC* amplification: $P_{MYCL} = 2.1 \times 10^{-5}$, $P_{MYCN} = 4.0 \times 10^{-3}$, $P_{none} = 0.12$).

replicates, *MYCL*-amplified tumors showed clear changes in gene transcription in comparison to tumors with only *TP53* deletion (**Supplementary Table 8**). For example, the progression-associated immunoglobulin basigin (*BSG*) was downregulated ($q = 0.029$) whereas fibroblast growth factor 1 (*FGF1*) was significantly upregulated ($q = 0.041$) in *MYCL*-mutant tumors (**Supplementary Fig. 7**). Amplification of *MYCL* was not associated with pretreatment PSA levels or clinical T category but was significantly associated with lower age at time of treatment ($P = 4.86 \times 10^{-3}$) and was significantly more frequent in tumors harboring *TMPRSS2-ERG* fusions (ERG^+; $P = 1.64 \times 10^{-2}$) (**Supplementary Fig. 8**).

We overexpressed *MYCL* (four separate isoforms) in two prostate cancer cell lines (22rv1 and LNCaP) and found substantial overlap in gene expression in these cells (relative to control-transfected cells) in comparison to expression in the primary tumors with *MYCL* amplification for which RNA was available (relative to tumors with copy-neutral *MYCL*; **Supplementary Fig. 9**). One such gene was *KLK3*, encoding PSA, the expression of which was significantly lower in cell lines overexpressing *MYCL* and in tumors harboring a *MYCL* amplification (*P* values ranging from 0.03 to 4.20×10^{-5} depending on the isoform). Moreover, we found that *MYCL* amplification was associated with increased proliferation in the absence of serum in both cell lines (data not shown).

Whole-genome sequencing of potentially curable prostate cancer

Recent studies have demonstrated genomic spatial heterogeneity within rapidly growing tumors[27,28], and prostate cancers with both an *SPOP* mutation and a *TMPRSS2-ERG* fusion can comprise two distinct foci (one with each aberration)[16]. To resolve the extent of intrafocal heterogeneity in localized, non-indolent prostate cancer, we performed extensive spatial sampling of five of the surgically resected prostates analyzed. We subjected formalin-fixed, paraffin-embedded preserved surgical tissue from these cases to exhaustive pathological study to identify additional disease foci using anatomical location, ERG and p63 expression. In total, we subjected DNA from 23 tumor regions in 5 patients to whole-genome sequencing, with 2–9 distinct foci from each patient (1 frozen specimen and 1–8 formalin-fixed, paraffin-embedded specimens per patient). We comprehensively analyzed SNVs, CNAs and genomic rearrangements (clinical annotation shown in **Supplementary Table 9**) and compared them to the interpatient heterogeneity described above. These specimens covered a broad range of genomic instability (PGA of 0.04 to 17.1%). Pathologically estimated tumor cellularity of >70% was confirmed by qpure analysis[29] (69–94%; **Supplementary Table 10**). To maximize potential translation into clinical practice (for example, using small amounts of tissue amenable to studies with biopsies), we generated whole-genome sequence data from low-input libraries of non-amplified genomic DNA (50 ng of input DNA) for all sequencing (**Supplementary Fig. 10**), with tumors at ~60× coverage (median coverage of 60.7×) and blood samples at ~40× coverage (median coverage of 44.2×; **Supplementary Table 11**).

Extensive structural heterogeneity

We initially focused on possible structural variations, which can be extensive in late-stage disease[21]. The index lesions harbored a median of 43 intrachromosomal rearrangements, 52 interchromosomal rearrangements and 18 inversions, along with 40 CNAs and a PGA of 0.33% (**Supplementary Fig. 11**). These numbers reflect higher genomic stability than in other solid tumors[30], including higher-risk prostate cancer[6]. There was extensive intertumoral diversity, with the total number of genomic rearrangements ranging from 20 to 197. All balanced genomic rearrangements are listed in **Supplementary Table 12**. We observed large differences between regions from the same prostate (**Fig. 3a**). For example, the index lesion of tumor CPCG0184 harbored 112 amplifications and 20 deletions comprising 1,262 genes (PGA = 1.71%). We analyzed four additional spatially separated tumor regions. These contained 1–177 regions of copy number gain and 2–194 regions of copy number loss (PGA = 0.2–3.7%). All 5 regions of CPCG0184 shared a small amplification on chromosome 7 (containing *LOC100506585*, *MIR595* and *PTPRN2*) and a large deletion on chromosome 8 containing 56 genes. By contrast, 2,144 genes had aberrant copy numbers in only one region.

These included the cancer-related genes *CCND1* (cyclin D1) and *CCNE2* (cyclin E2), 9 separate cytochrome P450 genes and the long noncoding RNA gene *HOTAIR*.

This extensive intrafocal heterogeneity was particularly evident at the level of genomic rearrangements (**Fig. 3b**). We performed a 'windowed' analysis to quantitatively assess the presence of genomic rearrangement hotspots, separating the genome into 3,113 bins of 1 Mb each (outside of chromosome ends). By assessing the frequency of rearrangement breakpoints within each bin, we identified numerous hotspots, including several on chromosome 1 that were altered

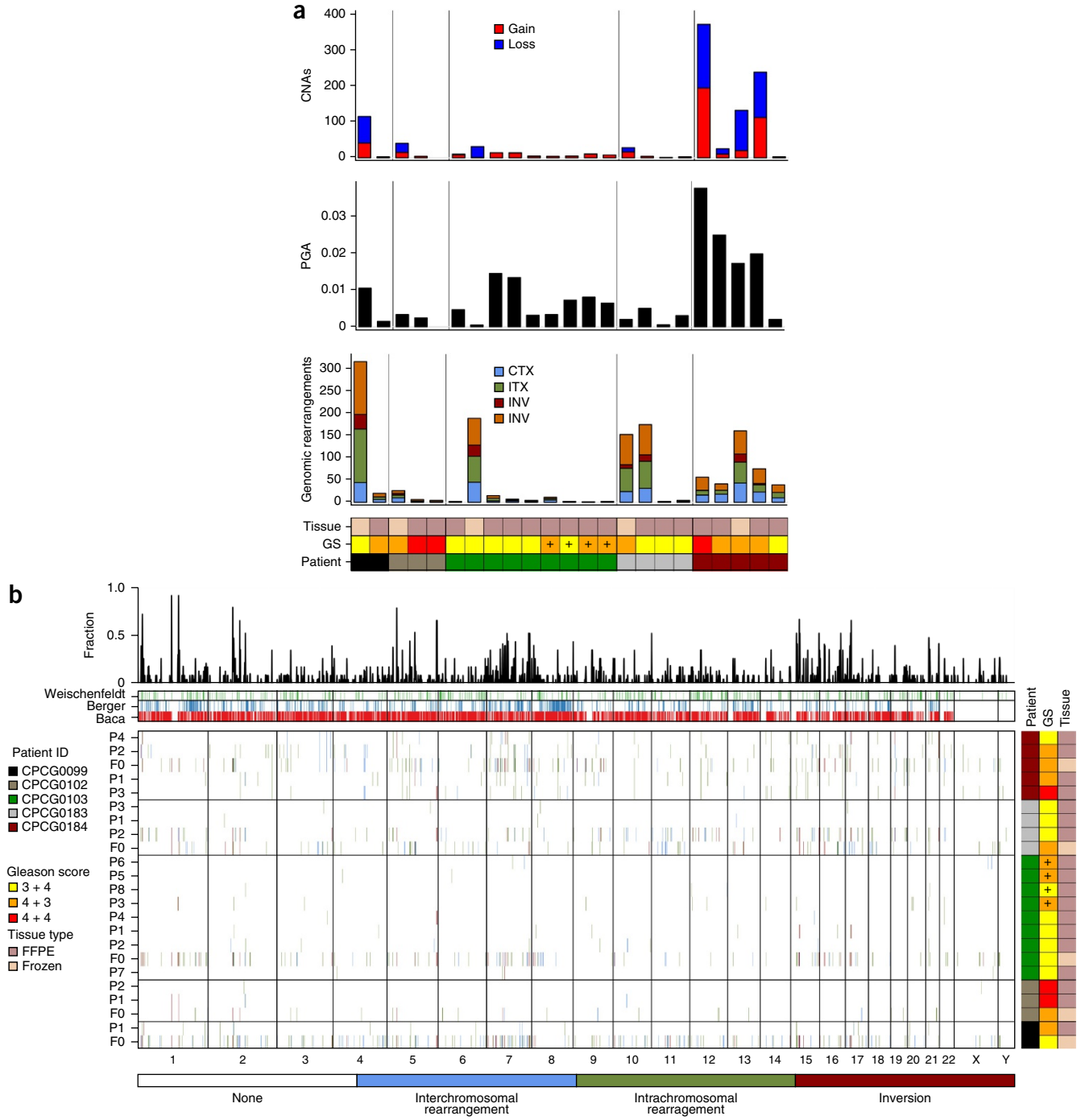

Figure 3 Structural variations in Gleason score 7 prostate cancer. (**a**) Structural genomic heterogeneity in prostate cancer. From top to bottom: CNA number for 23 prostate tumors, PGA and total number of genomic rearrangements per tumor stratified by type. Covariate colors are as for **Figure 2a**, with a plus sign in the Gleason score covariate indicating ERG$^+$ tumors by immunohistochemistry and red in the Gleason score covariate indicating Gleason score 4 + 4. Tissue samples were derived from frozen (light) or formalin-fixed, paraffin-embedded (FFPE; dark) sections. (**b**) Heterogeneity by breakpoints for balanced genomic rearrangements. From top to bottom: sliding-window analysis of breakpoint enrichment, breakpoints in the vicinity of those previously reported and heat map of breakpoint occurrence stratified by structural variation type. CTX, interchromosomal translocation; ITX, intrachromosomal translocation; INV, inversion; DEL, deletion. P numbers represent distinct tissue regions derived from formalin-fixed, paraffin-embedded samples, and F numbers represent fresh-frozen tissue.

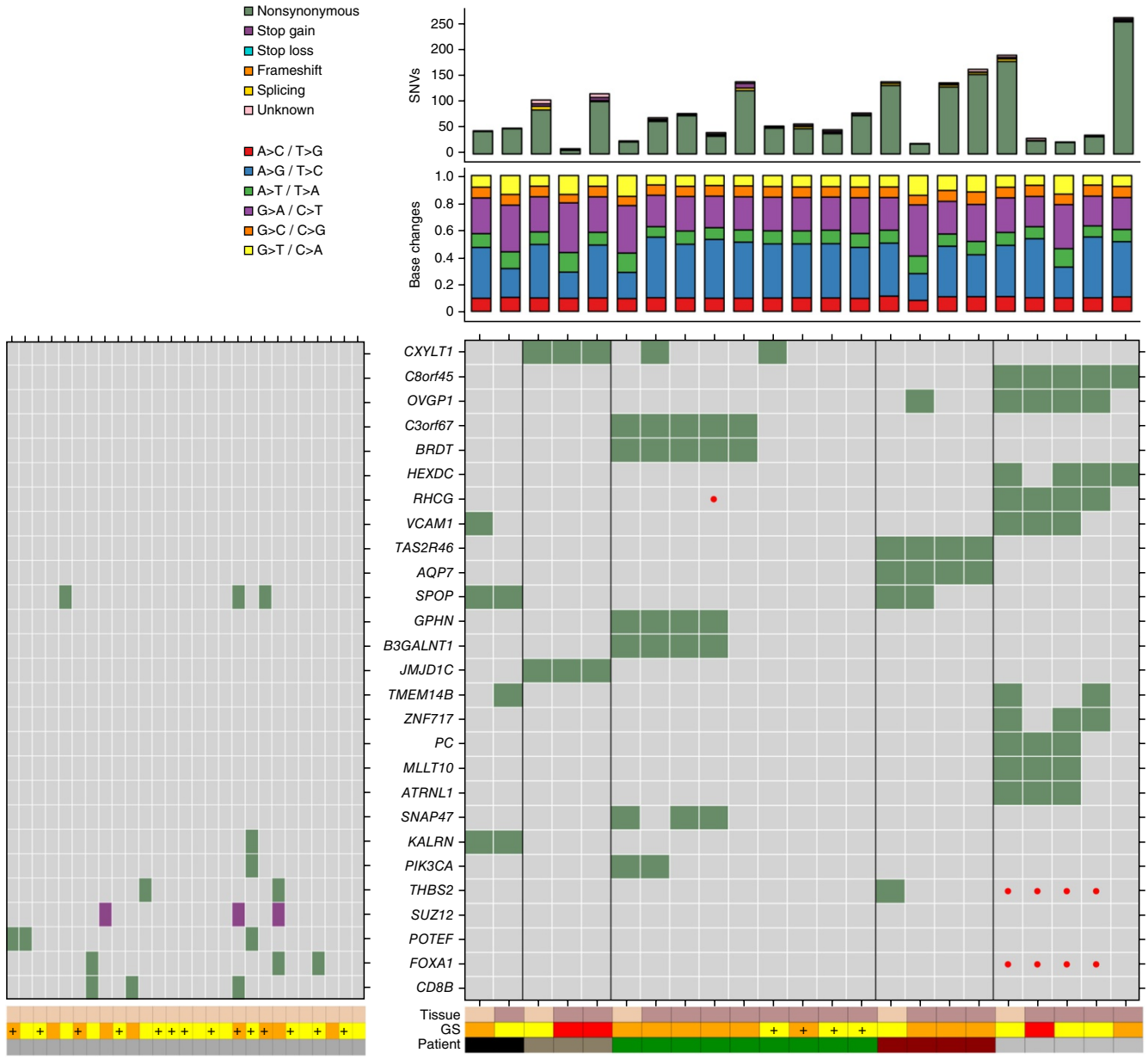

Figure 4 SNVs in Gleason score 7 prostate cancer. Integrated display of gene-level recurrent CNAs, functional single-nucleotide mutations (boxes: colored by function) and CNAs in those genes (dots: red, gain; blue, loss) across 23 samples from 5 tumors. A gene was included only if it had a validated functional (non-silent exonic or splice-site) SNV present in at least two samples, including at least one frozen sample. Bottom, tissue type (pink, frozen; mauve, formalin fixed, paraffin embedded), Gleason score (yellow, 3 + 3; orange, 3 + 4; dark orange, 4 + 3) and patient ID. Top, genome-wide base transition rates and all counts of functional SNV types. Left, mutations found in Gleason score 7 tumors in Baca *et al.*[21].

in every tumor region studied. Interestingly, all genome sequences derived from formalin-fixed, paraffin-embedded tumors showed a marked reduction in the number of genomic rearrangements (likely attributable to the smaller insert sizes for these libraries). This suggests that there are substantial false negative rates in sequencing such degraded DNA. We compared genomic rearrangements in our study to the genomic rearrangements detected in three publications of prostate cancer whole-genome sequencing[20,21,31]. There were no exact matches of breakpoints across any of the studies, suggesting a generalized genomic instability at the level of genomic rearrangement, akin to that outlined for copy number[10]. Of note, two recent prostate cancer sequencing studies[21,31] did not call genomic rearrangements on chromosome X or Y, making the current study the largest one thus far of genomic rearrangements on the sex chromosomes, on which we identified 35 total structural variations across the 5 tumors, including 1 inversion, 4 intrachromosomal rearrangements and 30 interchromosomal rearrangements (**Supplementary Table 12**).

Quiet point mutation profiles

We validated SNVs against genotyping microarray data (median of 91.3% accuracy; **Supplementary Fig. 12a**) and by deep resequencing of 1,743 SNVs detected in coding regions (**Supplementary Fig. 13**); SNVs that failed validation were excluded. The 5 index lesions harbored a median of 11,341 somatic SNVs (**Fig. 4**), with modest variation in the total number of somatic SNVs (9,375 to 15,512). By contrast, we observed dramatic differences in the number of protein-altering

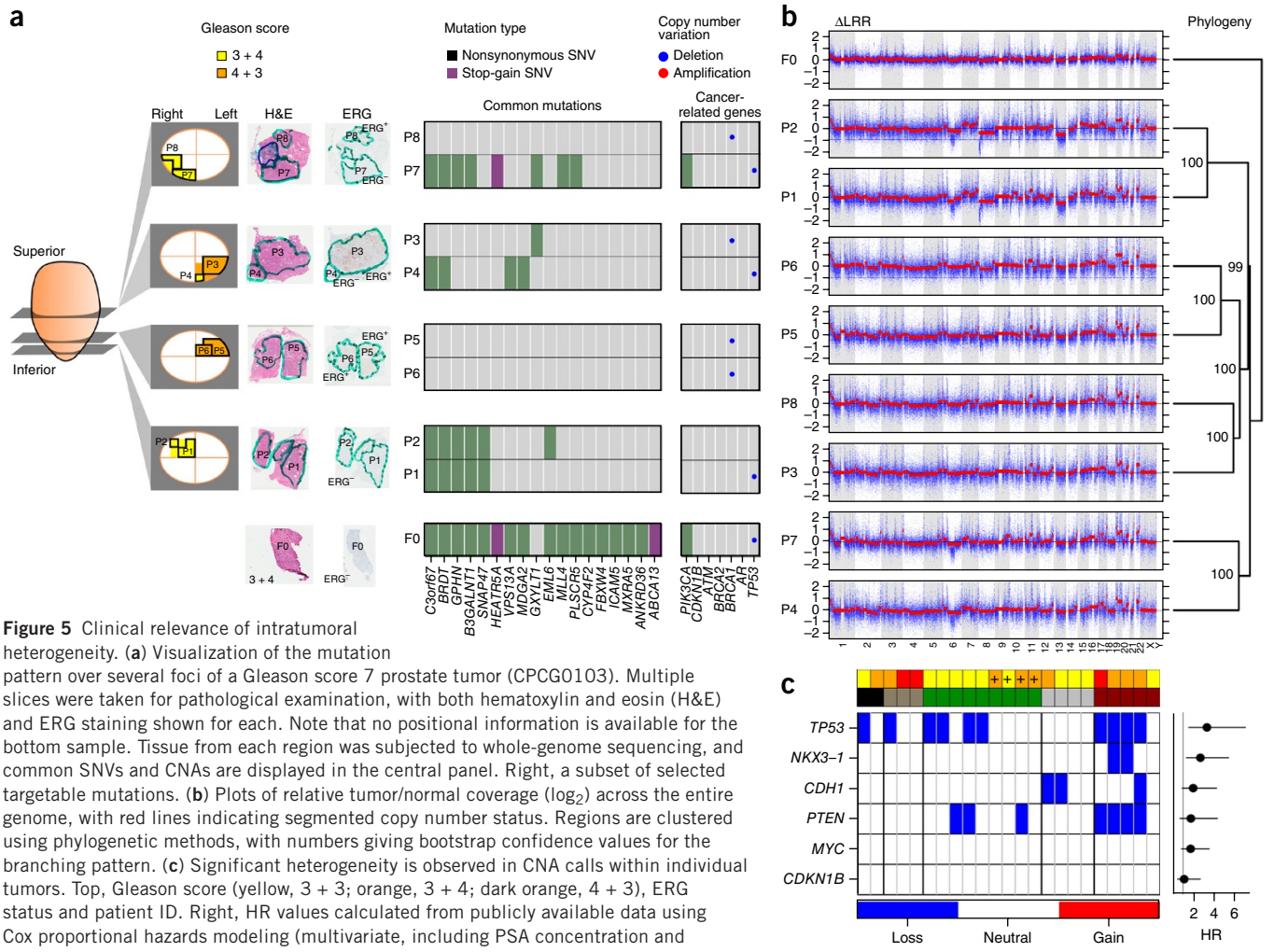

Figure 5 Clinical relevance of intratumoral heterogeneity. (**a**) Visualization of the mutation pattern over several foci of a Gleason score 7 prostate tumor (CPCG0103). Multiple slices were taken for pathological examination, with both hematoxylin and eosin (H&E) and ERG staining shown for each. Note that no positional information is available for the bottom sample. Tissue from each region was subjected to whole-genome sequencing, and common SNVs and CNAs are displayed in the central panel. Right, a subset of selected targetable mutations. (**b**) Plots of relative tumor/normal coverage (log_2) across the entire genome, with red lines indicating segmented copy number status. Regions are clustered using phylogenetic methods, with numbers giving bootstrap confidence values for the branching pattern. (**c**) Significant heterogeneity is observed in CNA calls within individual tumors. Top, Gleason score (yellow, 3 + 3; orange, 3 + 4; dark orange, 4 + 3), ERG status and patient ID. Right, HR values calculated from publicly available data using Cox proportional hazards modeling (multivariate, including PSA concentration and T stage), with 95% confidence intervals.

somatic SNVs (8 to 49). The SNV profiles of each tumor differed substantially, with only five genes exhibiting functional mutations in more than one tumor (four of which were identified in previous prostate cancer sequencing studies[15,16]). A recent study described *SPOP*, *FOXA1* and *MED12* as recurrently mutated genes in prostate cancer[16]. We observed nonsynonymous *SPOP* mutations in two patients (both ERG^-), but we detected no functional mutations in either *MED12* or *FOXA1*. Similarly, we did not observe mutations in *KMT2C* (also known as *MLL3*)[32], nor was the *AR* androgen receptor gene mutated or amplified in any samples.

We carefully validated our whole-genome sequencing data to ensure that formalin fixation and embedding in paraffin did not introduce artifacts relative to fresh-frozen tissues. Although somatic mutation rates in the fixed samples were elevated and showed a shift toward A>G/T>C transitions, we identified highly concordant sets of germline SNPs in each region (**Supplementary Fig. 14**), and median validation rates remained high (**Supplementary Fig. 12b**). To our knowledge, this represents the first report of whole-genome sequencing from routinely clinically achievable quantities (50 ng) of DNA from formalin-fixed, paraffin-embedded specimens.

We identified 26 genes mutated in at least 2 regions (**Fig. 4**) associated with extensive intrafocal heterogeneity. For example, *VCAM1* (vascular cell adhesion molecule 1) was mutated in three of five regions of CPCG0183. Similarly, *SNAP47* showed nonsynonymous point mutations in three of nine regions of CPCG0103. Although this study is underpowered, we did not observe statistically significant differences in mutation rates or specific mutations in the comparison of Gleason score 3 + 4 and 4 + 3 tumors at the SNV level, mirroring the well-powered findings at the CNA level outlined above. All functional coding SNVs identified are listed in **Supplementary Table 13**.

Clinical relevance of intraprostatic heterogeneity

We next investigated the potential clinical consequences of intrafocal heterogeneity using CPCG0103, a prostate with nine spatially distinct regions (**Fig. 5a**). Two Gleason score 3 + 4 regions (including the index lesion) exhibited mutations in *PIK3CA* encoding p.His1047Arg, the same mutation frequently found in breast, colon and other tumor types[33,34] and which predicts sensitivity to AKT inhibitors[35]. Thus, an analysis of the index lesion of this tumor might indicate sensitivity to AKT inhibition, yet seven of nine sequenced regions did not harbor this mutation (including all four Gleason score 4 + 3 regions). We conclude that one subclone, present exclusively in two regions of the prostate, has an actionable mutation in *PIK3CA*. Interestingly, these regions also exhibited loss of *TP53* (as did two others). Four regions from this individual showed loss of *BRCA1*, none of which had *TP53* loss. This heterogeneous mutational profile demonstrates the challenges in personalizing treatment for prostate cancer on the basis of

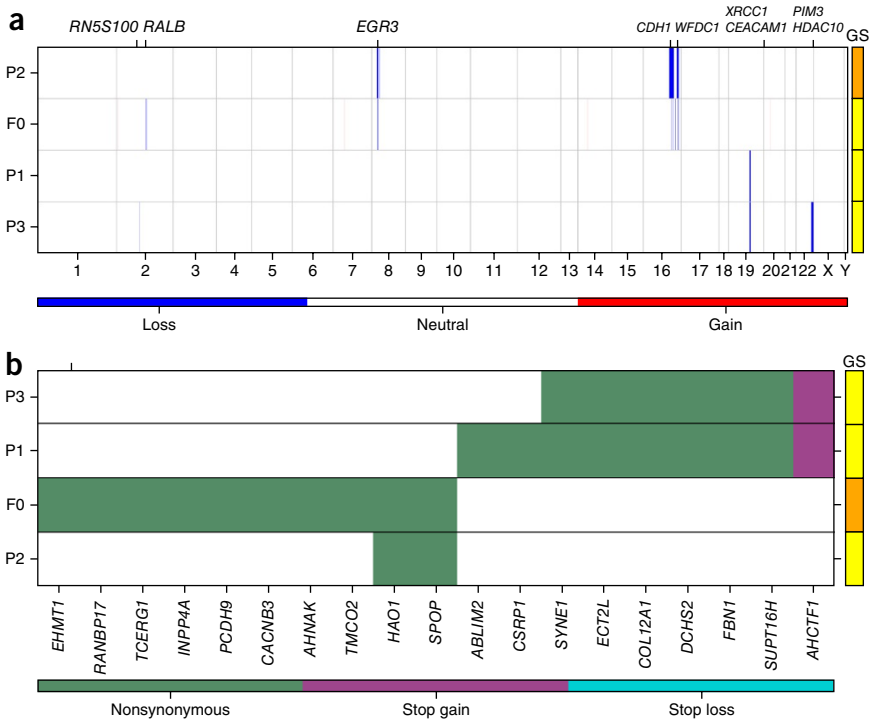

Figure 6 Evidence of multiclonal prostate cancer. (**a**,**b**) CPCG0183 exhibits two regions sharing no genomic aberrations at either the CNA (**a**) or SNV (**b**) level, suggestive of multiclonality.

a single focus; we made similar observations for the other four prostates (**Supplementary Figs. 15–18**).

We also observed intrafocal heterogeneity of structural variation: in the nine sequenced regions of CPCG0103, PGA ranged from 0.04 to 1.44% (~36-fold; similarly, the PGA in CPCG0184 varied by ~16-fold over the 4 regions analyzed). PGA is an independent prognostic factor, with every 1% increase in PGA associated with a 5–6% increase in hazard of biochemical failure[10]. These differences in intratumoral molecular profiles would change predictions of state-of-the-art prognostic biomarkers.

We exploited this large variability in structural variation to infer tumor phylogeny (outlined in **Supplementary Fig. 19**). We observed the same branched and complex phylogenetic relationships found in other tumor types[27,28] for CPCG0103 (**Fig. 5b**) and for the other four tumors studied (**Supplementary Figs. 20–23**). We examined the intratumoral heterogeneity in a set of candidate prognostic biomarkers for intermediate-risk prostate cancer (**Fig. 5c**), observing stark heterogeneity in the genomic profiles across tumor regions. For example, *NKX3-1*, a well-characterized tumor suppressor whose loss is highly prognostic in Gleason score 7 tumors (hazards ratio (HR) = 2.74; $P = 0.007$, Wald test), was deleted in two of five regions of CPCG0184. Thus, CNA-based biomarkers would yield different predictions of patient prognoses depending on which tumor region was analyzed.

Multiclonality of prostate cancer

The interfocal genomic diversity of prostate cancer was most evident in CPCG0183, from which four spatially separate regions were sequenced. This patient presented at age 64 years with elevated PSA levels (7.39 ng/ml). He was diagnosed with a Gleason score 6, T2a, ERG$^-$ adenocarcinoma. After radical prostatectomy, the patient was upstaged to a Gleason score 7 (4 + 3) tumor on the basis of the index lesion (CPCG0183-F0). The other three regions sequenced (CPCG0183-P1, CPCG0183-P2 and CPCG0183-P3) were ERG$^-$and Gleason score 3 + 4. All regions were highly cellular, with estimated tumor content from 67–80%.

The index lesion had a relatively stable genome, with a PGA of only 0.19%: 75 genes exhibited copy number loss (mostly in large deletions on chromosomes 8 and 16) and 13 genes showed copy number gain (including *MYCL*). The large deletions on chromosomes 8 and 16 were shared by the index lesion and the P2 region, although *MYCL* gain was not. By contrast, the other two tumor regions sequenced, P1 and P3, contained neither of these deletions. Instead, the region of chromosome 19 containing *XRCC1* was deleted in regions P1 and P3 but not in regions F0 and P2. Upon closer investigation, the F0/P2 and P1/P3 regions showed mutually exclusive CNA profiles (**Fig. 6a**).

Similarly, the index lesion contained ten nonsynonymous mutations, validated by either Sanger sequencing or deep resequencing studies (**Fig. 6b**). Of these, two were shared by the P2 region: *HAO1* and *SPOP* (chr. 17, g.47696425A>G; p.Phe133Ser). By contrast, *SPOP* mutations were completely absent in the P1 and P3 regions, which instead shared seven validated mutations, including gain of a stop codon in *AHCTF1*. We confirmed this pattern at the level of genomic rearrangements, where F0 and P2 showed extensive similarity, including inversion of *NOTCH2*, and no overlap with the P1 and P3 regions. These data strongly suggest that this individual harbored two separate prostate cancers, which have no underlying genetic etiology in common.

DISCUSSION

We performed the first systematic evaluation—at the level of whole-genome sequencing—of the genomic heterogeneity associated with localized, potentially curable multifocal prostate cancers, which were treated as non-indolent cancers. The vast majority of prostate cancer is diagnosed as localized disease[36]; however, many Gleason score 7 cancers are non-indolent and classified as intermediate-risk cancers requiring treatment. In up to one-third of these patients, local therapies fail to cure patients a priori[37]. Our studies provide insight into both the observed interpatient heterogeneity in clinical outcome and the potential complexity of intraglandular heterogeneity within an individual patient with prostate cancer. Previous studies of prostate cancer genomes have focused on aggressive and/or incurable cancers[15] or have evaluated a wide spectrum of clinical risk groups[16,31]. We previously observed that recurrent CNAs in genes such as *MYC*, *NKX3-1* and/or *PTEN* are heterogeneous and that genomic instability in low- and intermediate-risk prostate cancers is prognostic for biochemical recurrence following surgery or radiotherapy[10,37–39]. Recently, Cooper *et al.* described the genomic heterogeneity associated with atypical prostate cancers and adjacent non-malignant prostate epithelium[40]. Together, these studies represent a comprehensive assessment of the genomic heterogeneity associated with non-indolent, localized prostate cancer.

In assessing intraprostatic heterogeneity, we discovered and validated a new recurrent amplification of *MYCL*. Gain of *MYCL*

(but not of *MYC*) was nearly universally associated with *TP53* loss, suggesting cooperative dysregulation of these genes. Moreover, the transcriptome and unique biology of tumors harboring *MYCL* amplification (relative to those with *MYC* amplification) suggest that these tumors represent a new disease subtype. *MYCL* gains had not been observed in previous genome-wide studies, possibly owing to decreased probe densities on the older array or SNP chip assays used to generate CNA data. Our finding provides one of the first clear functional distinctions between MYC family members in human prostate cancer, although the functional relevance of *MYCL* amplification in prostate cancer remains to be fully defined. However, our data suggest that *MYCL* is associated with a unique CNA and gene expression profile in comparison to tumors harboring a *MYC* amplification, and overexpression of *MYCL* in two prostate cancer cell lines enhanced cell proliferation rates in serum-free medium (data not shown). *MYCL* amplification showed extensive intraprostatic heterogeneity by FISH analysis and was absent in adjacent tumor regions in two cases in which amplification was validated in the index lesion. Although much larger numbers of samples will be required to validate this hypothesis, this finding suggests that *MYCL* amplification may be preferentially localized to the index lesion. Moreover, the observation that *MYCL* amplification is inversely correlated with age at the time of treatment suggests that *MYCL* aberrations may potentially be a marker of earlier-onset prostate cancer; substantially larger data sets will be required to validate this hypothesis.

Although previous reports have hinted at multiclonality in prostate cancer[14,16], a robust evaluation of tumor phylogeny across a broad range of genomic aberrations has not been reported. We observed no shared CNAs and very few shared SNVs between disease foci, strongly suggesting the existence of multiclonal disease. This has important ramifications for the development of genetic prognostic and predictive biomarkers: (i) it must be clarified how specific individual malignant clones contribute to disease progression and (ii) biopsy-based diagnostic assays that miss genetically independent lesions could misclassify prostate cancer aggression and preclude optimal treatment with current therapies or newer targeted agents.

In conclusion, we believe our results provide a way forward to increase precision in prostate cancer prognosis and treatment. Although whole-genome and whole-exome sequencing have revolutionized the study of human tumor evolution and heterogeneity, these studies required microgram quantities of DNA for adequate sequencing complexity and depth. We have now shown that whole-genome sequencing can be completed on routine formalin-fixed, paraffin-embedded biopsies with an optimized low-input library protocol that allows for whole-genome sequencing (with at least 50× coverage). This may assist in the development of prognostic biomarkers based on sequencing of pretreatment materials, which could in turn be used to stratify patients before potentially curative treatments and/or direct patients to novel neoadjuvant or adjuvant therapies to prevent progression and prostate cancer lethality. Furthermore, although it is appreciated that diagnostic biopsy protocols can miss regions of more aggressive cancer (thereby understaging the patient), the genomic heterogeneity of pathologically identical regions of cancer must also be taken into consideration to achieve personalized prostate cancer medicine.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. mRNA microarray data are available at the Gene Expression Omnibus (GEO) under accession GSE64619; next-generation sequencing data are available at the European Genome-phenome Archive (EGA) under accession EGAS00001000549.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

The authors thank all members of the Boutros and Bristow laboratories for helpful suggestions. This study was conducted with the support of Movember funds through Prostate Cancer Canada and with the additional support of the Ontario Institute for Cancer Research, funded by the government of Ontario. This study was conducted with the support of the Ontario Institute for Cancer Research to P.C.B. through funding provided by the government of Ontario. This work has been funded by a Doctoral Fellowship from the Canadian Institutes of Health Research (CIHR) to E.L. The authors gratefully thank the Princess Margaret Cancer Centre Foundation and the Radiation Medicine Program Academic Enrichment Fund for support (to R.G.B.). R.G.B. is a recipient of a Canadian Cancer Society Research Scientist Award. This work was supported by Prostate Cancer Canada and is proudly funded by the Movember Foundation, grant RS2014-01. P.C.B. was supported by a Terry Fox Research Institute New Investigator Award and a CIHR New Investigator Award. This project was supported by Genome Canada through a Large-Scale Applied Project contract to P.C.B., S.P.S. and R. Morin.

AUTHOR CONTRIBUTIONS

Sample preparation and molecular biology: M.F., A. Meng, T.C., M.S., C.L.H., J.J., L.T., N.B., A.W., J.D.W., T.T.S., G.Z., A.D.P., A. Berlin, S.D.P. and A. Brown. Pathology analyses: D.T., B.T. and T.v.d.K. Statistics and bioinformatics: P.C.B., N.J.H., R.d.B., E.L., P.H.H.-Y., A. McPherson, V.Y.S., A.Z., N.S.F., J.L., Y.-J.S., J.W., T.A.B., T.T.S., C.P., F.N., X.L., K.C.C., J.S., M.A.C.-S.-Y., F.Y., R.E.D., L.C.C., G.M.C., E.J., M.H.W.S., H.C., S.K.G., J.H., A.D., M.P., C.F., F.H. and D.W. Initiation of the project: P.C.B., M.F., C.C., T.J.H., J.D.M., T.v.d.K., R.E., D.N. and R.G.B. Supervision of research: P.C.B., M.F., T.A.B., P.L., L.B.M., B.T., C.C.C., L.D.S., N.F., S.P.S., C.S., T.J.H., L.B.M., T.v.d.K. and R.G.B. Writing of the first draft of the manuscript: P.C.B. Writing and editing the revised manuscript: M.F., P.C.B. and R.G.B. All authors approved the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Mohler, J. *et al.* NCCN clinical practice guidelines in oncology: prostate cancer. *J. Natl. Compr. Canc. Netw.* **8**, 162–200 (2010).
2. D'Amico, A.V. *et al.* Cancer-specific mortality after surgery or radiation for patients with clinically localized prostate cancer managed during the prostate-specific antigen era. *J. Clin. Oncol.* **21**, 2163–2172 (2003).
3. Buyyounouski, M.K., Pickles, T., Kestin, L.L., Allison, R. & Williams, S.G. Validating the interval to biochemical failure for the identification of potentially lethal prostate cancer. *J. Clin. Oncol.* **30**, 1857–1863 (2012).
4. Villers, A., McNeal, J.E., Freiha, F.S. & Stamey, T.A. Multiple cancers in the prostate. Morphologic features of clinically recognized versus incidental tumors. *Cancer* **70**, 2313–2318 (1992).
5. Nichol, A.M., Warde, P. & Bristow, R.G. Optimal treatment of intermediate-risk prostate carcinoma with radiotherapy: clinical and translational issues. *Cancer* **104**, 891–905 (2005).
6. Taylor, B.S. *et al.* Integrative genomic profiling of human prostate cancer. *Cancer Cell* **18**, 11–22 (2010).
7. Lapointe, J. *et al.* Genomic profiling reveals alternative genetic pathways of prostate tumorigenesis. *Cancer Res.* **67**, 8504–8510 (2007).
8. Paris, P.L. *et al.* Whole genome scanning identifies genotypes associated with recurrence and metastasis in prostate tumors. *Hum. Mol. Genet.* **13**, 1303–1313 (2004).
9. Penney, K.L. *et al.* mRNA expression signature of Gleason grade predicts lethal prostate cancer. *J. Clin. Oncol.* **29**, 2391–2396 (2011).
10. Lalonde, E. *et al.* Tumour genomic and microenvironmental heterogeneity for integrated prediction of 5-year biochemical recurrence of prostate cancer: a retrospective cohort study. *Lancet Oncol.* **15**, 1521–1532 (2014).
11. Olmos, D. *et al.* Prognostic value of blood mRNA expression signatures in castration-resistant prostate cancer: a prospective, two-stage study. *Lancet Oncol.* **13**, 1114–1124 (2012).
12. Cortese, R. *et al.* Epigenetic markers of prostate cancer in plasma circulating DNA. *Hum. Mol. Genet.* **21**, 3619–3631 (2012).
13. Ruijter, E.T., van de Kaa, C.A., Schalken, J.A., Debruyne, F.M. & Ruiter, D.J. Histological grade heterogeneity in multifocal prostate cancer. Biological and clinical implications. *J. Pathol.* **180**, 295–299 (1996).

14. Lindberg, J. *et al.* Exome sequencing of prostate cancer supports the hypothesis of independent tumour origins. *Eur. Urol.* **63**, 347–353 (2013).
15. Grasso, C.S. *et al.* The mutational landscape of lethal castration-resistant prostate cancer. *Nature* **487**, 239–243 (2012).
16. Barbieri, C.E. *et al.* Exome sequencing identifies recurrent *SPOP*, *FOXA1* and *MED12* mutations in prostate cancer. *Nat. Genet.* **44**, 685–689 (2012).
17. Ren, S. *et al.* RNA-seq analysis of prostate cancer in the Chinese population identifies recurrent gene fusions, cancer-associated long noncoding RNAs and aberrant alternative splicings. *Cell Res.* **22**, 806–821 (2012).
18. Prensner, J.R. *et al.* Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. *Nat. Biotechnol.* **29**, 742–749 (2011).
19. Kumar, A. *et al.* Exome sequencing identifies a spectrum of mutation frequencies in advanced and lethal prostate cancers. *Proc. Natl. Acad. Sci. USA* **108**, 17087–17092 (2011).
20. Weischenfeldt, J. *et al.* Integrative genomic analyses reveal an androgen-driven somatic alteration landscape in early-onset prostate cancer. *Cancer Cell* **23**, 159–170 (2013).
21. Baca, S.C. *et al.* Punctuated evolution of prostate cancer genomes. *Cell* **153**, 666–677 (2013).
22. Zhou, Z. *et al.* Synergy of p53 and Rb deficiency in a conditional mouse model for metastatic prostate cancer. *Cancer Res.* **66**, 7889–7898 (2006).
23. Edwards, J., Krishna, N.S., Witton, C.J. & Bartlett, J.M. Gene amplifications associated with the development of hormone-resistant prostate cancer. *Clin. Cancer Res.* **9**, 5271–5281 (2003).
24. Pugh, T.J. *et al.* The genetic landscape of high-risk neuroblastoma. *Nat. Genet.* **45**, 279–284 (2013).
25. Rushlow, D.E. *et al.* Characterisation of retinoblastomas without *RB1* mutations: genomic, gene expression, and clinical studies. *Lancet Oncol.* **14**, 327–334 (2013).
26. Penn, L.J., Brooks, M.W., Laufer, E.M. & Land, H. Negative autoregulation of c-Myc transcription. *EMBO J.* **9**, 1113–1121 (1990).
27. Gerlinger, M. *et al.* Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N. Engl. J. Med.* **366**, 883–892 (2012).
28. Bashashati, A. *et al.* Distinct evolutionary trajectories of primary high-grade serous ovarian cancers revealed through spatial mutational profiling. *J. Pathol.* **231**, 21–34 (2013).
29. Song, S. *et al.* qpure: a tool to estimate tumor cellularity from genome-wide single-nucleotide polymorphism profiles. *PLoS ONE* **7**, e45835 (2012).
30. Beroukhim, R. *et al.* The landscape of somatic copy-number alteration across human cancers. *Nature* **463**, 899–905 (2010).
31. Berger, M.F. *et al.* The genomic complexity of primary human prostate cancer. *Nature* **470**, 214–220 (2011).
32. Lindberg, J. *et al.* The mitochondrial and autosomal mutation landscapes of prostate cancer. *Eur. Urol.* **63**, 702–708 (2013).
33. Samuels, Y. *et al.* High frequency of mutations of the *PIK3CA* gene in human cancers. *Science* **304**, 554 (2004).
34. Janku, F. *et al. PIK3CA* mutation H1047R is associated with response to PI3K/AKT/mTOR signaling pathway inhibitors in early-phase clinical trials. *Cancer Res.* **73**, 276–284 (2013).
35. Sangai, T. *et al.* Biomarkers of response to Akt inhibitor MK-2206 in breast cancer. *Clin. Cancer Res.* **18**, 5816–5828 (2012).
36. Djulbegovic, M. *et al.* Screening for prostate cancer: systematic review and meta-analysis of randomised controlled trials. *BMJ* **341**, c4543 (2010).
37. Zafarana, G. *et al.* Copy number alterations of *c-MYC* and *PTEN* are prognostic factors for relapse after prostate cancer radiotherapy. *Cancer* **118**, 4053–4062 (2012).
38. Locke, J.A. *et al. NKX3.1* haploinsufficiency is prognostic for prostate cancer relapse following surgery or image-guided radiotherapy. *Clin. Cancer Res.* **18**, 308–316 (2012).
39. Locke, J.A. *et al.* Allelic loss of the loci containing the androgen synthesis gene, *StAR*, is prognostic for relapse in intermediate-risk prostate cancer. *Prostate* **72**, 1295–1305 (2012).
40. Cooper, C.S. *et al.* Analysis of the genetic phylogeny of multifocal prostate cancer identifies multiple independent clonal expansions in neoplastic and morphologically normal prostate tissue. *Nat. Genet.* **47**, 367–372 (2015).

ONLINE METHODS

Samples. All patients underwent either image-guided radiotherapy (IGRT) or radical prostatectomy for intermediate-risk prostate cancer, as defined by the NCCN (Gleason score = 6, PSA concentration >10 and <20 ng/ml, cT1 or T2; Gleason score = 7, PSA concentration <20, cT1 or T2). The collection of fresh-frozen pretreatment research biopsies from patients receiving radiotherapy has been described previously[41]. Fresh-frozen radical prostatectomy specimens were obtained from the University Health Network BioBank. Formalin-fixed, paraffin-embedded tissue blocks were obtained from the Department of Pathology, University Health Network. Whole blood was collected, and informed consent, consistent with local research ethics board (REB) and International Cancer Genome Consortium (ICGC) guidelines, was obtained at the time of clinical follow-up. Previously collected tumor tissue was used according to University Health Network REB-approved study protocols (UHN 06-0822-CE and UHN 11-0024-CE).

To confirm Gleason score and tumor cellularity, all tumor specimens were independently evaluated by three genitourinary pathologists (T.v.d.K., B.T. and D.T.) using scanned slides stained with hematoxylin and eosin. Formalin-fixed, paraffin-embedded tumor regions were located in the prostate and selected for macrodissection according to gross description and ERG overexpression status (immunohistochemistry; clone 9FY, Biocare Medical). Median follow-up time was calculated on the basis of surviving patients.

Sample processing. Selected samples were cut into sections of 60 × 10 µm, with a 4-µm section for hematoxylin and eosin staining generated every ten cuts and alternate sections taken for DNA and RNA extraction. Sections stained with hematoxylin and eosin were marked by a genitourinary pathologist (T.v.d.K. or D.T.) to indicate areas suitable for macrodissection (with >70% tumor cellularity). Manual macrodissection was performed using sterile scalpel blades, and DNA was obtained by phenol-chloroform extraction, as previously reported[41]. DNA was extracted from whole blood using an ArchivePure DNA Blood kit (5 PRIME) at the Applied Molecular Profiling Laboratory at the Princess Margaret Cancer Centre. All DNA samples were quantified using a Qubit 2.0 Fluorometer (Life Technologies). For frozen biopsy samples from patients undergoing IGRT, 100 ng of genomic DNA was used as a template for whole-genome amplification with the GenomePlex Complete WGA2 kit (Sigma-Aldrich). Whole genome–amplified DNA was used for OncoScan SNP microarrays; all whole-genome sequencing was carried out using non-amplified genomic DNA.

SNP microarrays. SNP microarrays were performed using 200 ng of whole genome–amplified (IGRT biopsy) or genomic (radical prostatectomy) DNA on Affymetrix OncoScan FFPE Express 2.0 SNP arrays. We confirmed that whole-genome amplification did not markedly affect CNA and SNP profiles by comparing genomic and whole genome–amplified DNA samples from three independent specimens (**Supplementary Table 14**). We similarly compared duplicate genomic and whole genome–amplified DNA samples to evaluate interassay variability (data not shown). We analyzed SNP probe assays from Affymetrix to call CNAs using BioDiscovery Nexus Copy Number software with default parameters. The data from Affymetrix assays were processed in batches, and in some cases liftOver was used to map aberrations from the genome reference hg18 to the hg19 reference sequence. When the liftOver process deleted a portion of the CNA, the CNA was removed from the analysis. We then used the CNAs detected in tumor samples to identify genes with altered copy number using GENCODE (v17) reference gene annotation[42]. PGA was calculated for each sample by dividing the number of base pairs that were involved in a copy number change in each sample by the total length of the genome.

To estimate the cellularity and purity of our tumor samples, we used the qpure algorithm[29]. Notably, qpure requires the log R ratio (LRR) and B allele frequency (BAF) for SNP array probes. These values were computed for the OncoScan array platform using the two intensity values provided for each probe, corresponding to the hybridization of two alleles, using the following equations: $\mathrm{LRR} = \log_2 (X + Y)$ and $\mathrm{BAF} = Y/(X + Y)$, where Y and X are intensity values corresponding to the minor and major alleles, respectively. We used qpure to compute the cellularity of our samples with default parameters and selected the output (tumorpurity.mixture.gam.adjust) as our cellularity estimate.

Given a set of copy number profiles for each sample, as well as copy number intensities for all the probes on our SNP array platform, we used GISTIC2.0 to study the recurrence of gene copy number variations in our sample set[43]. As input, GISTIC2.0 requires a file for each tumor sample that contains the average copy number intensities for each segmented region along the chromosomes. Therefore, for each sample, we created a 'somatic' copy number profile along the chromosomes. We refer to any copy number change as somatic if we did not observe the same event in the corresponding blood sample. Each profile segmented the chromosomes into regions with neutral events, copy number loss or copy number gain. For each segmented region, we computed the corresponding copy number intensity by averaging the copy number intensities of the probes in that segment obtained from the SNP array. We used this input file along with the SNP array probe assay reference file to run a GISTIC2.0 MATLAB script with default parameters (-genegistic 1 -smallmem 1 -broad 1 -brlen 0.5 -conf 0.90, $q = 0.25$). We used GISTIC2.0 output to identify significantly recurrent gene amplifications and deletions and to study these genes in more detail. Specifically, we identified gene copy number changes in all 75 samples in the cohort of fresh-frozen tumors (**Fig. 1a**). Statistical comparisons between PGA and the number of genes with CNAs in Gleason score 3 + 4 and 4 + 3 tumors employed the Wilcoxon rank-sum test, as implemented in R (v3.0.1). Comparisons of CNA number and size between tumor groups were performed using the Wilcoxon rank-sum test. Specific genes showing different CNA frequencies were identified using Fisher's exact test. Six tumors with mutations in both *MYCL* and *MYCN* and one tumor with mutations in both *MYC* and *MYCL* were excluded from intergroup comparisons to avoid confounding.

Quantitative PCR. Amplifications of *MYCL* and *MYC* were verified using TaqMan Copy Number assays (Life Technologies), according to the manufacturer's instructions and using *RPPH1* (RNase P) as a copy-neutral gene for comparison; we verified the absence of *RPPH1* CNAs in our OncoScan data sets. The following probes were used: *MYCL*, Hs02582452_cn; *MYC*, Hs02758348_cn.

To verify the size of the *MYCL* amplification, we performed qPCR using ten custom-designed probes flanking the *MYCL* gene, encompassing an approximately 2-Mb region (1 Mb on each side of the gene). The relative positions of these probes are shown in **Supplementary Figure 5**.

FISH analysis. FISH analysis was conducted as previously described[37]. *MYCL* was detected using probe RP1-118J21 (chr. 1: 40,225,143–40,391,639). *CEP1* probe (encompassing chr. 1: 119,400,971–119,941,816) was from Agilent Technologies. FISH conditions were optimized on metaphase spreads of human lymphocytes, as previously described[37]. The tissues used were as follows: four benign prostates, taken from men who underwent cystoprostatectomy and had no pathological evidence of prostate cancer, and six prostate cancer specimens, using formalin-fixed, paraffin-embedded tissue taken from the dominant lesion of disease, immediately adjacent to the frozen tissue used in the corresponding molecular analysis. Four of the six prostate cancer specimens had PCR-validated *MYCL* amplification, whereas the remaining two were PCR validated as copy neutral for *MYCL*.

Co-occurrence of alterations for *MYC* family members. To calculate the probability of observing no overlap between *MYC* family CNAs, we simulated (n = 1,000,000) the observed mutation rates of the 3 *MYC* family isoforms under the null distribution of independence and applied the resulting proportions as estimates of the probability of observing n overlaps.

RNA sample preparation. For RNA extractions, ice-cold tumor tissue maintained on RNase-free microscope slides was macrodissected from 10-µm sections and transferred to an Eppendorf tube on ice. Tissue was digested in lysis buffer, and total RNA was extracted using the mirVana miRNA Isolation kit (Life Technologies), according to the manufacturer's instrucitons.

mRNA array analysis. For the analysis of expression data, a total of 24 frozen prostate cancer tissue samples were available. These comprised ten samples that were assayed using the HuGene 2.0 Affymetrix array (Centre for Applied Genomics; batch 1). The remaining 14 samples (batch 2) were assayed using

doi:10.1038/ng.3315

the Human Transcriptome Array (HTA), also from Affymetrix, processed at the London Regional Genomics Centre (Ontario, Canada). CNA calls were made using the OncoScan array on the same tissue samples.

Quantified intensity files (CEL files) for each array were loaded into the R statistical environment (v3.0.2). Files generated for batch 1 (HuGene2.0 arrays) were loaded and normalized using the robust multichip average (RMA) algorithm available in the oligo package for R (v1.26.2), using the default platform design information (pd.hugene.2.0.st v3.8.0). Similarly, files generated for batch 2 (HTA) were loaded and normalized by RMA using the affy package (v1.40.0) with a custom chip definition file (CDF; hta20hsentrezgcdf v18.0.0)[44]. Each batch was examined for spatial and distributional homogeneity; no outliers were identified.

Normalized batch 1 transcript clusters were mapped to Entrez Gene IDs and gene symbols using hugene20sttranscriptcluster.db (v2.12.1), available from Bioconductor. As these arrays provide coverage of both coding and noncoding transcripts, not all probe sets were annotated; unannotated probe sets were labeled as 'other' and remain uniquely identifiable using the probe set IDs.

After mapping of probes to gene annotations, we took the mean signal intensity where multiple probes mapped to the same gene and included only genes present on both arrays. To correct for batch effects, we used the ComBat algorithm as implemented in the sva R package (v3.8.0). Before and after heat maps (**Supplementary Fig. 24**) highlight the successful removal of batch effects. Copy number measures for *MYC*, *MYCL*, *MYCN* and *TP53* were used as covariates in the empirical Bayes model.

Pairwise *t* tests were carried out to calculate *P* values and fold changes for all mRNA abundances between groups, where groups were defined by CNA status. *P* values were corrected for multiple testing using the false discover rate (FDR) method implemented in the base R function, p.adjust. After identification of significant genes, gene ontology analysis was carried out using GOEAST (Gene Ontology Enrichment Analysis Software Toolkit). The raw text data are available as **Supplementary Table 15**. All analyses were carried out using R (v3.0.2). Tumors with aberrations in multiple *MYC* family genes were excluded from statistical analyses.

Cell line mRNA analyses. All preprocessing was performed using R (v3.0.3). Background correction and normalization algorithms were implemented in the affy package (v1.40.0) for the Bioconductor (v2.14) open source project[45]. The RMA algorithm was applied to the raw intensity data[46]. Quality control was conducted using R (v3.0.3). Density and heat map plots with default parameterization were created using lattice (v0.20-29). Interarray correlation was calculated using the stats package. No outliers were detected in the LNCaP and 22RV1 cell lines. Data annotation was performed in R (v3.0.3) via the hta20hsentrezg.db (v18.0.0) package. Probe IDs were mapped to both Entrez Gene IDs and HUGO Gene Nomenclature Committee gene symbols. Unsupervised machine learning was performed in R (v3.0.3). DIANA was used as the clustering method, and 1-Pearson's correlation was used as the distance metric between columns and between rows (genes and samples, respectively). Heat maps were generated using lattice (v0.20-29). Statistical analysis was performed in R (v3.1.1) using the limma package[47] (v3.20.9) of the Bioconductor (v2.14) open source project[45]. After preprocessing, we modeled mRNA expression in each cellular compartment as a univariate linear model of effects from differential experimental conditions (untreated, overexpression of the complete coding sequence of *MYCL*, overexpression of transcript variant 1 of *MYCL*, overexpression of transcript variant 2 of *MYCL*, overexpression of transcript variant 3 of *MYCL*). All model-based *t* tests were corrected using an empirical Bayes moderation of standard error followed by an FDR adjustment for multiple testing. Model-based *t* tests coupled with Bayesian moderation of the standard error were implemented in the limma package. FDR was used to adjust for multiple testing using the function p.adjust from the core R stats package (limma and *t*-test analysis). Genes showing significant differential expression were filtered on the basis of q value < 0.1 and absolute $\log_2$ |fold change| > 1. Once significant genes were identified in the cell line samples, we then compared these genes with those associated with *MYCL*-amplified primary tumors. All comparisons were made in R (v3.1.2). The top 19 genes were selected on the basis of q value < 0.1 and absolute $\log_2$ |fold change| >1.5 in both cell lines (22RV1 and LNCaP). These genes were then mapped to the genes in the primary tumor samples. We next conducted an unpaired two-tailed *t* test on these genes from *MYCL*-amplified and *MYCL*–wild type primary tumors. LNCaP cells were obtained from the American Type Culture Collection. 22rv1 cells were a generous gift from Y. Pinthus (McMaster University). The absence of mycoplasma contamination was routinely confirmed by Hoechst 33258 staining and/or PCR.

Genome sequencing. Qubit-quantified genomic DNA (50 ng; non-amplified) was sheared to 300-bp fragments using the Covaris S2 Ultrasonicator, and 3× volume AMPure XP SPRI bead (Beckman Coulter Genomics, A63881) clean-up was performed. The bead-DNA mixture was transferred to a 96-well PCR plate (Eppendorf, 0030133404) for the remainder of library construction and all subsequent SPRI bead clean-up steps. Libraries were constructed using enzymatic reagents from KAPA Library Preparation kits (KAPA Biosciences, KK8201) according to protocols for end repair, A-tailing and adaptor ligation[48]. Adaptor-ligated libraries were enriched using optimized PCR conditions by adding 3 µl of 25 µM Illumina forward and revers paired-end enrichment primers (Integrated DNA Technologies), 75 µl of 2× KAPA HiFi HotStart ReadyMix (KAPA Biosciences, KK2602) and 33 µl of nuclease-free water (Life Technologies, AM993) to 36 µl of eluted DNA and amplified across three individual PCR tubes. Libraries were incubated in a Verti 96-well Thermal Cyclers (Life Technologies) for 45 s at 98 °C and cycled ten times for 15 s at 98 °C, 30 s at 65 °C and 30 s at 72 °C. After a 0.6× SPRI bead clean-up step, post-PCR enriched libraries were eluted in 40 µl of elution buffer (Qiagen, 19086) and validated using the Agilent Bioanalyzer High-Sensitivity DNA kit (Agilent Technologies, 5067-4626).

Libraries were quantified on the Illumina Eco Real-Time PCR instrument using KAPA Illumina Library Quantification kits (KAPA Biosciences, KK4835) according to the standard protocols from the manufacturer. Paired-end sequencing of 2 × 101 cycles was carried out for all libraries on the Illumina HiSeq 2000 platform, and samples were sequenced with the number of lanes predicted to yield an uncollapsed coverage of 50× and 30× for tumor and normal samples, respectively.

Sequence alignment and variant calling. Next-generation sequence data were processed using several third-party and custom tools. To automate the analysis, SeqWare[49] workflows were used for initial data processing, and Perl and R scripts were developed for downstream automation. Files with raw base call and intensity data were transferred from the Illumina HiSeq 2000 sequencing instruments to the network storage system during the course of the sequencing run. These files were converted to FASTQ format using Illumina's CASAVA (v1.8.2) software and managed using a SeqWare workflow. The FASTQ files were then aligned to the UCSC GRCh37/hg19 human reference (with no repeat masking) using Novocraft's Novoalign short-read aligner (v2.07.14). Reads that aligned to multiple locations in the reference genome were retained and recorded in the aligned data file to a maximum of five alignment positions. Novoalign produced output in SAM format (v1.4) with properly configured read groups generated by Picard (v1.56). Reads were converted to BAM format, sorted by coordinates and indexed using Picard (v1.56). All BAM files generated from a single library were merged, and PCR duplicate reads for each library were removed using Picard (v1.56). The final library BAM files were then merged to a single sample and tissue BAM file using Picard (v1.56). SAMtools (v0.1.18) was used to filter out unaligned reads (SAMFlag = 4) and reads aligning to multiple locations (mapping quality < 30)[50].

SNVs were identified using the Genome Analysis Toolkit (GATK; v1.3.16)[51,52] as outlined in GATK Best-Practices for Variant Detection v2. BAM files for both blood and tumor samples were analyzed together for each sample. For tumors with multiple sampled formalin-fixed, paraffin-embedded regions, each tumor region was analyzed with its corresponding blood sample as a pair. Each sample BAM file was locally realigned around known insertion and deletion events as identified in the dbSNP 135 reference[53]. The locally realigned BAM files were then processed for base quality recalibration to adjust the error profile to represent the alignment error profile. SNVs were identified for each matched sample (consisting of both the tumor and blood locally realigned and quality-recalibrated BAM files) using the UnifiedGenotyper walker within GATK, which generated a Variant Call Format (VCF) file. The VCF file was then compressed and indexed using Tabix[54]. To parallelize

the variant calling pipeline, each chromosome in the GRCh37/hg19 reference was used to create regions for GATK indel realignment, recalibrating base qualities and calling SNVs.

Somatic mutations were identified using the SNV file containing both tumor and blood variants for each matched set of samples in the VCF file. A custom tool was developed to identify base call differences between the tumor and blood samples at each VCF file locus, and resulting variants were identified as somatic mutations and annotated in the final VCF file. Bases that were common to the tumor and blood samples but differed from the UCSC GRCh37/hg19 reference were identified as germline mutations.

SNV filtering and identification of recurrent somatic SNVs. After somatic SNV and indel calling, identified variants were passed through an annotation pipeline. Variants were functionally annotated by ANNOVAR (v2012-10-23)[55], using the RefGene database. Nonsynonymous, stop-loss, stop-gain, frameshift and splice-site SNVs (based on RefGene annotations) were considered functional. Variants were filtered using the Perl implementation of tabix (v0.2.6)[54], removing variants found in any of the following databases: dbSNP137 (modified to remove somatic and clinical variants, with variants with the following flags excluded: SAO = 2/3, PM, CDA, TPA, MUT and OM), 1000 Genomes Project (v3), Complete Genomics 69 whole genomes, NHLBI exome sequencing study (Exome Variant Server, NHLBI GO Exome Sequencing Project; accessed March 2013), duplicate gene database (v68)[56], ENCODE DAC and Duke Mapability Consensus Excludable databases (comprising poorly mapping reads, repeat regions, and mitochondrial and ribosomal DNA)[57] and the Fuentes database of likely false positive variants[58]. Variants were whitelisted (and retained, independently of the presence on other filters) if they were contained within the Catalogue of Somatic Mutations in Cancer (COSMIC) database (v62)[59]. Plots of recurrence and functional annotations were created in the R statistical environment (v2.15.3) using the lattice (v0.20-15), latticeExtra (v0.6-24) and RColorBrewer (v1.0-5) packages.

The mutation rate per megabase was calculated by dividing the number of somatic point mutations after validation by the count of callable loci × 10^{-6}. A callable locus was defined as a genomic position where the aligned read depth in both the tumor and corresponding normal samples was eight or greater. This count was calculated for each filtered, deduplicated tumor-normal pair of BAM files using the SAMtools (v0.1.18) depth utility and straightforward use of the standard Linux command line utilities awk (v1.3.3) and wc (v8.5) within a custom Perl wrapper script.

Data from a previous study of prostate cancer genomics were incorporated to provide a baseline for our analysis[21]. Somatic SNVs in exons were obtained from Supplementary Table S3A (file mmc3_Somatic_DNA_alterations.xlsx). Variants were passed through the same in-house processing pipeline for annotation and filtering via blacklists for potential germline variants (dbSNP137 and 1000 Genomes Project) or false positives (Fuentes, duplicate gene, in-house validation failures) as used for our own data. Variants appearing in COSMIC v82 were whitelisted. Variants were reannotated using ANNOVAR with the RefGene hg19 database. Functional variants are displayed in **Figure 4**.

Detection of genomic rearrangements. Sequenced genomes were preprocessed using the same pipeline for quality control and alignment as for variant calling. To detect genomic rearrangements, we used deStruct (v.0.1.0; A. McPherson, C.S. and S.P.S., unpublished data), a modification of nFuse[60]. In particular, deStruct uses high-throughput DNA sequencing data without requiring RNA input, in contrast to nFuse. With deStruct, we are able to obtain breakpoints and calls for rearrangements such as deletions, insertions, inversions, and intra- and interchromosomal translocations. Intrachromosomal translocations herein are defined as two breakpoints from a single chromosome that have been joined, as described by others[61]. In this version of deStruct, rearrangements with lengths from 1,000 bp to megabases can be detected. Analysis of calls recurring across samples was performed in part using BEDTools (v2.17.0)[62], as well as R (v3.0.1), with the lattice package (v0.20-15) for visualization.

Characterization of the median tumor. For structural rearrangements (genomic rearrangements, CNAs and PGA) and the total number of somatic SNVs, the median was the median of the values for all tumors. For specific SNV classes, the fraction of each tumor's SNVs that were classified as exonic by ANNOVAR was calculated. The median of these proportions was then used to calculate the expected number of exonic SNVs by multiplying by the median count of somatic SNVs. A similar procedure was used to determine the expected number of nonsynonymous and other functional mutations in the median tumor.

Phylogenetic dendrogram construction. The entire process for dendrogram creation was performed in the R statistical environment (v2.15.2), using the ape (v3.0-6) and phangorn (v1.7-1) packages. For each SNV data dendrogram, a preliminary tree was created using neighbor joining, which was used as a starting point to run maximum-likelihood analysis. The resulting unrooted tree was rooted using the patient's blood reference sample as an outlier. Finally, the external branches were extended to artificially force the tree to be ultrametric (a requirement of a dendrogram). Dendrograms for CNA data were created using neighbor joining with pairwise distances defined as the proportion of CNA values (no change, addition or deletion) that were different for the two samples. As with the SNV data, the blood reference was used as an outlier to root the tree and external branch lengths were artificially extended to force the tree into dendrogram form. Bootstrapping was performed on phylogenetic trees with more than three leaves. At each iteration of the bootstrapping procedure, the data points within the CNA profile were sampled with replacement until there was the same number of data points as for the original CNA profiles. Each resampled set of data points was used to create a tree by neighbor joining with the distance between each pair of CNA profiles being the percentage of the CNAs that did not match for the two profiles. A total of 100,000 bootstrap iterations were used for each tree, and the bootstrap values on each branch point are the percentage of times out of the 100,000 repetitions that that the branch point was in the tree. Bootstrapping was carried out in R (v3.0.3) using the phangorn (v1.99-7) and ape (v3.1-1) packages.

Deep sequencing–based SNV validation. The overall SNV matrix was processed to filter variants and rank somatic mutations on the basis of their recurrence in the 28 tumor-normal matched samples. A union of the predicted somatic mutations was then used for validation on an orthogonal platform. The final somatic mutation list included 1,037 SNVs that were annotated as functional across all 23 tumor regions analyzed, including 951 nonsynonymous, 27 stop-gain and 5 stop-loss mutations, 29 of unknown function and 25 in splicing regions. We added to these ~1,000 mutations seen in other prostate cancer whole-genome sequencing data. Predicted somatic mutations were prepared for validation using a custom Life Technologies AmpliSeq panel (AmpliSeq Primer Design Pipeline v2.0) and sequenced using the Life Technologies Ion Torrent Personal Genome Machine (PGM). Primers for the custom panel were designed using the AmpliSeq online primer design pipeline (v2.0). The primer design pipeline was able to generate 1,614 primer pairs, encompassing 1,743 mutations, to cover 10 blood, 9 frozen tumor and 15 formalin-fixed, paraffin-embedded tumor samples in total (4 samples had insufficient material to perform validation). The amplicon size was designed to be 150 bp as recommended in the AmpliSeq documentation for formalin-fixed, paraffin-embedded samples. To identify both false positive and false negative somatic mutations, the primers for the 1,743 mutations were applied to the blood and prostate tumor samples.

Library preparation and sequencing were performed by EdgeBio. A read depth of 500× was targeted across all 1,614 amplicons using Ion Torrent PGM 318 chips. At the completion of the sequencing run, FASTQ files were provided by EdgeBio and downloaded using their web portal. A custom alignment reference was created based on each specifically targeted amplicon with a flanking region of ±10 bp. The FASTQ files were then aligned to the target reference using Novoalign (v3.00.03), sorted by coordinates and converted to BAM format using Picard (v1.90). Read groups were added to the BAM files using Picard. A modified pileup file was generated using the predicted SNVs and the amplicon-aligned BAM file to identify the base counts at each SNV position.

A statistical model was developed to classify and filter mutations for each tumor-normal matched sample. A χ^2 test between each tumor and normal sample was applied to determine a P value at each SNV. To compensate for

doi:10.1038/ng.3315

multiple testing, the Bonferroni correction method was applied to the *P* values. We then calculated the Euclidean distance between the vectors of proportions between the tumor and matched normal samples. To account for uneven coverage across the samples, a standard deviation based on the total sample base count at each SNV position and the mean base coverage across all positions was calculated. For normal tissues, a ternary allele proportion was calculated and analyzed to quantify potential false positive variants in blood samples. Mutations were classified as somatic if the adjusted *P* value was less than 0.25, the Euclidean distance was greater than 0.15, coverage was no more than 1 s.d. from the mean of all samples and the ternary allele proportion for normal tissue was less than 0.05.

PCR and Sanger sequencing verification. Genomic DNA (10 ng) was amplified using the primer pairs shown in **Supplementary Table 16**. The presence of a single PCR product was verified by electrophoresis on a 1.5% agarose gel, and 50 ng of PCR product was sequenced from both the 5′ and 3′ ends by Sanger sequencing using the original PCR primers. Trace files (AB1 files) were converted to a multi-FASTA format using phredPhrap from the CONSED (v17.0) suite of tools[63]. Each line in the FASTA file represented a single amplicon sequence. These sequences were then aligned to the UCSC GRCh37/hg19 human reference using the bwa-sw algorithm from BWA (v0.7.0)[64] in SAM format. The SAM file was converted to BAM format, sorted by coordinate order and indexed using SAMtools (v0.1.18)[50]. An inspection of each SNV position of interest was performed using the SAMtools pileup algorithm. For sequences with base calls showing differences between the normal and tumor samples, manual inspection of the trace files was performed using FinchTV (v1.3.1; Geospiza), and a call was made for the given position.

Visualization. All visualizations were generated in the R statistical environment (v3.0.1) using the lattice (v0.20-15), latticeExtra (v0.6-24) and VennDiagram (v1.6.4)[65] packages, along with pdfTeX (v3.1415926-1.40.10). Schematics were created in Inkscape (v0.48) for Ubuntu.

41. Ishkanian, A.S. *et al.* High-resolution array CGH identifies novel regions of genomic alteration in intermediate-risk prostate cancer. *Prostate* **69**, 1091–1100 (2009).
42. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012).
43. Mermel, C.H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* **12**, R41 (2011).
44. Dai, M. *et al.* Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res.* **33**, e175 (2005).
45. Gentleman, R.C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **5**, R80 (2004).
46. Irizarry, R.A. *et al.* Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res.* **31**, e15 (2003).
47. Smyth, G.K. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.* **3**, Article3 (2004).
48. Fisher, S. *et al.* A scalable, fully automated process for construction of sequence-ready human exome targeted capture libraries. *Genome Biol.* **12**, R1 (2011).
49. O'Connor, B.D., Merriman, B. & Nelson, S.F. SeqWare Query Engine: storing and searching sequence data in the cloud. *BMC Bioinformatics* **11** (suppl. 12), S2 (2010).
50. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
51. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
52. DePristo, M.A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
53. NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **41**, D8–D20 (2013).
54. Li, H. Tabix: fast retrieval of sequence features from generic TAB-delimited files. *Bioinformatics* **27**, 718–719 (2011).
55. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
56. Ouedraogo, M. *et al.* The duplicated genes database: identification and functional annotation of co-localised duplicated genes across genomes. *PLoS ONE* **7**, e50653 (2012).
57. Gerstein, M.B. *et al.* Architecture of the human regulatory network derived from ENCODE data. *Nature* **489**, 91–100 (2012).
58. Fuentes Fajardo, K.V. *et al.* Detecting false-positive signals in exome sequencing. *Hum. Mutat.* **33**, 609–613 (2012).
59. Forbes, S.A. *et al.* COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res.* **39**, D945–D950 (2011).
60. McPherson, A. *et al.* nFuse: discovery of complex genomic rearrangements in cancer using high-throughput sequencing. *Genome Res.* **22**, 2250–2261 (2012).
61. Wang, J. *et al.* CREST maps somatic structural variation in cancer genomes with base-pair resolution. *Nat. Methods* **8**, 652–654 (2011).
62. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
63. Ewing, B., Hillier, L., Wendl, M.C. & Green, P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* **8**, 175–185 (1998).
64. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
65. Chen, H. & Boutros, P.C. VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. *BMC Bioinformatics* **12**, 35 (2011).